Proyecto Integrador de Ciencia de Datos

⚠️ Instrucciones generales

Para este proyecto integrador, cada estudiante debe **seleccionar un modelo** de aprendizaje supervisado entre los siguientes:

- **Regresión lineal**
- **Regresión logística**
- **Clasificación** KNN
- **Clustering**

🔷 *Se han agregado los scripts de ejemplo para cada uno de estos modelos al siguiente repositorio de GitHub: https://github.com/MauroKrdna/UdeColombia.git*

Allí encontrarán notebooks explicativos para regresión lineal, regresión logística y clasificación, que servirán como guía para su correcta implementación.

💻 ¿Cómo clonar el repositorio en tu PC?

Para trabajar localmente con los archivos, puedes clonar el repositorio usando los siguientes pasos:

1. Abre una terminal (o Git Bash en Windows).
2. Ubícate en la carpeta donde deseas guardar el proyecto.
3. Ejecuta el siguiente comando:

```
git clone https://github.com/MauroKrdna/UdeColombia.git
```

4. Una vez clonado, abre la carpeta del proyecto con Jupyter Notebook, Google Colab o tu entorno de desarrollo preferido.

🎯 Objetivo General

Aplicar una metodología de análisis predictivo para resolver un problema realista con datos estructurados, implementando desde la carga y preparación de datos hasta la evaluación del modelo y la realización de predicciones.

Bases de datos disponibles:

Los estudiantes pueden elegir una de las siguientes bases de datos propuestas o seleccionar una base de datos propia, siempre y cuando contenga variables suficientes para aplicar un modelo de regresión o clasificación y una variable objetivo clara.

1. **Student Performance Dataset**
   Predecir el *Performance Index* de los estudiantes a partir de factores como horas de estudio, puntajes anteriores, sueño, etc.
2. **Graduate Admissions Dataset**
   Estimar la *Chance of Admit* en un posgrado según variables como GRE, TOEFL, GPA, experiencia en investigación, etc.
3. **Cereal Dataset**
   Clasificar o predecir variables como tipo de cereal, fabricante o calificación (*rating*) a partir de sus componentes nutricionales.

El proyecto debe incluir:

1. **Carga y preparación de datos**
   - Limpieza de datos
   - Codificación de variables si es necesario (por ejemplo, categóricas a numéricas)
2. **División del dataset**
   - Separación en conjunto de entrenamiento y prueba (ej: 80% / 20%)
   - Se debe utilizar una semilla fija (`random_state`) para garantizar la reproducibilidad del modelo.
3. **Entrenamiento del modelo**
   - Implementación del modelo seleccionado: regresión lineal, regresión logística o clasificación.
   - Entrenamiento con los datos preparados.
4. **Evaluación del modelo**
   - Para **regresión lineal**, calcular:
     - `R² Score` (coeficiente de determinación)
     - `MSE` (Mean Squared Error)
   - Para **regresión logística** o **clasificación**, calcular:
     - `Accuracy`
     - `Matriz de confusión`
5. **Realizar predicciones**
   - Mostrar **al menos tres predicciones individuales** realizadas por el modelo.
   - Explicar claramente los valores de entrada que se usaron para cada predicción.

Entrega esperada

El trabajo debe entregarse en un **notebook** (Jupyter, Google Colab o similar) y contener:

✅ Código paso a paso
✅ C**onclusiones**, que debe incluir:

- Interpretación clara de los valores obtenidos en la evaluación del modelo (R^2, MSE, Accuracy, matriz de confusión).
- Discusión sobre las predicciones realizadas.

💬 Sugerencia final

Analiza primero si la variable objetivo que vas a predecir es **numérica continua** (ideal para regresión), o **categórica** (ideal para clasificación o regresión logística), antes de decidir tu modelo.